Technical Assessment: SKY

Author: Matthew W. Noble, DPhil [About the Author]

Date of Submission: 2018 · January · 15

Preamble

This technical assessment was performed by Matthew W. Noble for the Data Scientist role at SKY. The deadline for submission was 07:00 on 2018/01/15. The task is outlined below:

Exploratory and Predictive Data Analysis

For this analysis you will use the road safety data available from here:

http://data.gov.uk/dataset/road-accidents-safety-data

Please use latest 5 years (or more) STATS19 data **(accident, causalities and vehicles tables) for 2012 to 2016.**

Business Understanding

Description

These files provide detailed road safety data about the circumstances of personal injury road accidents in GB, the types (including Make and Model) of vehicles involved and the consequential casualties. The statistics relate only to personal injury accidents on public roads that are reported to the police, and subsequently recorded, using the STATS19 accident reporting form.

Task

The purpose of this analysis is:

- To summarize the main characteristics of the data, and obtain interesting facts that are worth highlighting.
- Identity and quantify associations (if any) between the number of causalities (in the Accidents table) and other variables in the data set.
- Explore whether it is possible to predict accident hotspots based on the data.

Your data analysis should consist of the following components:

1. A sort description and justification of the steps taken.

2. Visualization and description of the data.

3. Performance and evaluations results of the predictive model.

4. Insights gained from the analysis.

Please send a copy of your completed analysis to HR within a week (7 days).

Data Understanding

The data.gov.uk website has "ADDITIONAL LINKS" at the bottom of the page. These include:

- Road Safety - Blood Alcohol Content Data Guide
  - http://data.dft.gov.uk.s3.amazonaws.com/road-accidents-safety-data/BloodAlcoholContentDataGuide.xls

- Road Safety - Digital Breath Test Data Guide
  - http://data.dft.gov.uk.s3.amazonaws.com/road-accidents-safety-data/DigitalBreathTestDataGuide.xls

- Published statistics and supporting documents
  - https://www.gov.uk/government/collections/road-accidents-and-safety-statistics

- Lookup up tables for variables
  - http://data.dft.gov.uk/road-accidents-safety-data/Road-Accident-Safety-Data-Guide.xls

- Brief guide to road accidents and safety data
  - http://data.dft.gov.uk/road-accidents-safety-data/Brief-guide-to%20road-accidents-and-safety-data.doc

Of particular importance were the "Lookup tables for the variables" and the:

Brief guide to road accidents and safety data: Great Britain

These files provide detailed data about the circumstances of personal injury road accidents in Great Britain from 2005 onwards, the types of vehicles involved and the consequential casualties. The statistics relate only to personal injury accidents on public roads that are reported to the police, and subsequently recorded, using the STATS19 accident reporting form. Information on damage-only accidents, with no human casualties or accidents on private roads or car parks are not included in this data.

Very few, if any, fatal accidents do not become known to the police although it is known that a considerable proportion of non-fatal injury accidents are not reported to the police. Figures for deaths refer to persons killed immediately or who died within 30 days of the accident. This is the usual international definition, adopted by the Vienna Convention in 1968.

A list of the variables contained in the files can be found at:

http://data.dft.gov.uk/road-accidents-safety-data/Road-Accident-Safety-Data-Guide.xls

As well as giving details of date, time and location, the accident file gives a summary of all reported vehicles and pedestrians involved in road accidents and the total number of casualties, by severity. Details in the casualty and vehicle files can be linked to the relevant accident by the "Accident_Index" field. The Longitude and Latitude data is based on WGS 1984.

Further information on the data collection process, including the form (STATS19) used to collect the statistics, the instructions for it's completion (STATS20) and the definitions used in these data are available on the DfT website at:

http://www.dft.gov.uk/statistics/series/road-accidents-and-safety/

A copy of the form (STATS19) used to collect the data is available on the Department's website at:

http://assets.dft.gov.uk/statistics/series/road-accidents-and-safety/stats19-road-accident-injury-statistics-report-form.pdf

A copy of the guidance and instructions for completion of the form is available at:

http://assets.dft.gov.uk/statistics/series/road-accidents-and-safety/stats20-instructions-for-the-completion-of-road-accident-report-form-stats19-2011.pdf

Data Preparation

The Data tables were downloaded and as .zip files, they were unpacked and saved as .csv files and inspected. (The code has an option to create the DataFrames directly from the .gov website, but I assumed that I would be working offline). The files Accidents, Casualties, and Vehicles were all tables of individual integers or strings. An additional lookup .xlsx file provided the correct substitutions. These were made and the data generally cleaned up. There were mistakes and errors in the data and these were dealt with on an ad-hoc basis.

Data Tables and their Column Meanings

Accident Circumstances

- Accident Index
  - Gives a unique index for each accident and links to Vehicle and Casualty data.

- Police Force
  - Indicates which one of the 51 police constabulary's handled the accident.

- Accident Severity
  - Denotes if the accident was Fatal, Serious, or Slight.

- Number of Vehicles
  - Indicates the number of vehicles involved as an integer.

- Number of Casualties
  - Indicates the number of casualties involved as an integer.

- Date (DD/MM/YYYY)
  - Indicates the date in British calendar format.

- Day of Week
    - Lookup for Monday through Sunday.

- Time (HH:MM)
    - Indicates the time of the accident.

- Location Easting OSGR (Null if not known)
    - Map information.

- Location Northing OSGR (Null if not known)
    - Map information.

- Longitude (Null if not known)
    - Map information.

- Latitude (Null if not known)
    - Map information.

- Local Authority (District)
    - Indicates which one of the 416 UK districts where the accident occurred.

- Local Authority (Highway Authority - ONS code)
    - Indicates which one of the 207 UK highway authorities where the accident occurred.

- 1st Road Class
    - Denotes the road class as being Motorway, A(M), A, B, C, or Unclassified

- 1st Road Number
    - Indicates the 1st road number as an integer.

- Road Type
    - Denotes the Road Type as being Roundabout, One Way Street, Dual Carriageway, Single Carriageway, Slip Road, Unknown, One Way Street / Slip Road, or Data Missing or Out of Range.

- Speed limit
    - Indicates the speed limit of the road.

- Junction Detail
    - Denotes the junction as Not at junction or within 20 metres, Roundabout, Mini-roundabout, T or staggered junction, Slip road, Crossroads, More than 4 arms (not roundabout), Private drive or entrance, Other junction, or Data Missing or Out of Range.

- Junction Control
    - Indicates the control at the junction as Not at junction or within 20 metres, Authorised person, Auto traffic signal, Stop sign, Give way or uncontrolled, Data missing or out of range.

- 2nd Road Class

    - Denotes the 2nd road class as Not at junction or within 20 metres, Motorway, A(M), A, B, C, or Unclassified.

- 2nd Road Number
    - Indicates the 2nd road number as an integer.

- Pedestrian Crossing-Human Control
    - Indicates the type of human controlled pedestrian as None within 50 metres, Control by school crossing patrol, Control by other authorised person, or Data missing or out of range.

- Pedestrian Crossing-Physical Facilities
    - Indicates the type of physical pedestrian crossing as No physical crossing facilities within 50 metres, Zebra, Pelican, puffin, toucan or similar non-junction pedestrian light crossing, Pedestrian phase at traffic signal junction, Footbridge or subway, Central refuge, or Data missing or out of range.

- Light Conditions
    - Indicates the light conditions as Daylight, Darkness - lights lit, Darkness - lights unlit, Darkness - no lighting, Darkness - lighting unknown, or Data missing or out of range.

- Weather Conditions
    - Indicates the weather conditions as Fine no high winds, Raining no high winds, Snowing no high winds, Fine + high winds, Raining + high winds, Snowing + high winds, Fog or mist, Other, Unknown, or Data missing or out of range.

- Road Surface Conditions
    - Indicates the road surface conditions as Dry, Wet or damp, Snow, Frost or ice, Flood over 3cm. deep, Oil or diesel, Mud, or Data missing or out of range.

- Special Conditions at Site
    - Indicates any special conditions at the site as None, Auto traffic signal - out, Auto signal part defective, Road sign or marking defective or obscured, Roadworks, Road surface defective, Oil or diesel, Mud, or Data missing or out of range.

- Carriageway Hazards
    - Indicates carriageway hazards as None, Vehicle load on road, Other object on road, Previous accident, Dog on road, Other animal on road, Pedestrian in carriageway - not injured, Any animal in carriageway (except ridden horse), or Data missing or out of range.

- Urban or Rural Area
    - Indicates if the area is Urban, Rural, or Unallocated.

- Did Police Officer Attend Scene of Accident
    - Indicates if a police officer attended as Yes, No, or No - accident was reported using a self completion form (self rep only).

- Lower Super Ouput Area of Accident_Location (England & Wales only)

Casualties

- Accident Index
  - Gives a unique index for each accident and links to Vehicle and Casualty data.

- Vehicle Reference
  - Links the Casualties data table with the Vehicles data table

- Casualty Reference
  - It’s not made explicitly clear what this refers to.

- Casualty Class
  - Indicates the casualty class as Driver or rider, Passenger, or Pedestrian.

- Sex of Casualty
  - Indicates the sex of the casualty as Male, Female, or Data missing or out of range.

- Age of Casualty [Only exists for 2016, 2015, & 2014]
  - Indicates the age of the casualty as an integer.

- Age Band of Casualty
  - Indicates the age band of the casualty as 0 - 5, 6 - 10, 11 - 15, 16 - 20, 21 - 25, 26 - 35, 36 - 45, 46 - 55, 56 - 65, 66 - 75, Over 75, or Data missing or out of range.

- Casualty Severity
  - Indicates the severity of the casualty as Fatal, Serious, or Slight.

- Pedestrian Location
  - Indicates the pedestrian location as Not a Pedestrian, Crossing on pedestrian crossing facility, Crossing in zig-zag approach lines, Crossing in zig-zag exit lines, Crossing elsewhere within 50m. of pedestrian crossing, In carriageway, crossing elsewhere, On footway or verge, On refuge, central island or central reservation, In centre of carriageway - not on refuge, island or central reservation, In carriageway, not crossing, Unknown or other, Data missing or out of range.

- Pedestrian Movement
  - Indicates the pedestrian movement as Not a Pedestrian, Crossing from driver's nearside, Crossing from nearside - masked by parked or stationary vehicle, Crossing from driver's offside, Crossing from offside - masked by parked or stationary vehicle, In carriageway, stationary - not crossing (standing or playing), In carriageway, stationary - not crossing (standing or playing) - masked by parked or stationary vehicle, Walking along in carriageway, facing traffic, Walking along in carriageway, back to traffic, Unknown or other, or Data missing or out of range.

- Car Passenger
  - Indicates the car passenger as Not car passenger, Front seat passenger, Rear seat passenger, or Data missing or out of range.

- Bus or Coach Passenger

  - Indicates the bus or coach passenger as Not a bus or coach passenger, Boarding, Alighting, Standing passenger, Seated passenger, or Data missing or out of range.

- Pedestrian Road Maintenance Worker (From 2011)
  - Indicates the pedestrian road maintenance worker as No / Not applicable, Yes, Not Known, or Data missing or out of range.

- Casualty Type
  - Indicate the casualty type as Pedestrian, Cyclist, Motorcycle 50cc and under rider or passenger, Motorcycle 125cc and under rider or passenger, Motorcycle over 125cc and up to 500cc rider or passenger, Motorcycle over 500cc rider or passenger, Taxi/Private hire car occupant, Car occupant, Minibus (8 - 16 passenger seats) occupant, Bus or coach occupant (17 or more pass seats), Horse rider, Agricultural vehicle occupant, Tram occupant, Van / Goods vehicle (3.5 tonnes mgw or under) occupant, Goods vehicle (over 3.5t. and under 7.5t.) occupant, Goods vehicle (7.5 tonnes mgw and over) occupant, Mobility scooter rider, Electric motorcycle rider or passenger, Other vehicle occupant, Motorcycle - unknown cc rider or passenger, or Goods vehicle (unknown weight) occupant.

- Casualty IMD Decile [Only exists for 2016, & 2015]
  - Indicate the casualty IMD decile as Most deprived 10%, More deprived 10-20%, More deprived 20-30%, More deprived 30-40%, More deprived 40-50%, Less deprived 40-50%, Less deprived 30-40%, Less deprived 20-30%, Less deprived 10-20%, Least deprived 10%, or Data missing or out of range.

- Casualty Home Area Type
  - Indicate the casualty home area type as Urban area, Small town, Rural, or Data missing or out of range.

Vehicles

- Accident Index
  - Gives a unique index for each accident and links to Vehicle and Casualty data.

- Vehicle Reference
  - Links the Casualties data table with the Vehicles data table

- Vehicle Type
  - Indicates the vehicle type as Pedal cycle, Motorcycle 50cc and under, Motorcycle 125cc and under, Motorcycle over 125cc and up to 500cc, Motorcycle over 500cc, Taxi/Private hire car, Car, Minibus (8 - 16 passenger seats), Bus or coach (17 or more pass seats), Ridden horse, Agricultural vehicle, Tram, Van / Goods 3.5 tonnes mgw or under, Goods over 3.5t. and under 7.5t, Goods 7.5 tonnes mgw and over, Mobility scooter, Electric motorcycle, Other vehicle, Motorcycle - unknown cc, Goods vehicle - unknown weight, or Data missing or out of range.

- Towing and Articulation
  - Indicates any towing or articulation as No tow/articulation, Articulated vehicle, Double or multiple trailer, Caravan, Single trailer, Other tow, or Data missing or out of range.

- Vehicle Manoeuvre
    - Indicates the vehicle manoeuvre as Reversing, Parked, Waiting to go - held up, Slowing or stopping, Moving off, U-turn, Turning left, Waiting to turn left, Turning right, Waiting to turn right, Changing lane to left, Changing lane to right, Overtaking moving vehicle - offside, Overtaking static vehicle - offside, Overtaking - nearside, Going ahead left-hand bend, Going ahead right-hand bend, Going ahead other, or Data missing or out of range.

- Vehicle Location-Restricted Lane
    - Indicates the vehicle location as On main c'way - not in restricted lane, Tram/Light rail track, Bus lane, Busway (including guided busway), Cycle lane (on main carriageway), Cycleway or shared use footway (not part of main carriageway), On lay-by or hard shoulder, Entering lay-by or hard shoulder, Leaving lay-by or hard shoulder, Footway (pavement), Not on carriageway, or Data missing or out of range.

- Junction Location
    - Indicates the junction location as Not at or within 20 metres of junction, Approaching junction or waiting/parked at junction approach, Cleared junction or waiting/parked at junction exit, Leaving roundabout, Entering roundabout, Leaving main road, Entering main road, Entering from slip road, Mid Junction - on roundabout or on main road, or Data missing or out of range.

- Skidding and Overturning
    - Indicates the skidding or overturning as None, Skidded, Skidded and overturned, Jack-knifed, Jack-knifed and overturned, Overturned, or Data missing or out of range.

- Hit Object in Carriageway
    - Indicates if an object was hit in the carriageway as None, Previous accident, Road works, Parked vehicle, Bridge (roof), Bridge (side), Bollard or refuge, Open door of vehicle, Central island of roundabout, Kerb, Other object, Any animal (except ridden horse), or Data missing or out of range.

- Vehicle Leaving Carriageway
    - Indicates the vehicle leaving the carriageway as Did not leave carriageway, Nearside, Nearside and rebounded, Straight ahead at junction, Offside on to central reservation, Offside on to central res + rebounded, Offside - crossed central reservation, Offside, Offside and rebounded, or Data missing or out of range.

- Hit Object off Carriageway
    - Indicates if an object was hit off the carriageway as None, Road sign or traffic signal, Lamp post, Telegraph or electricity pole, Tree, Bus stop or bus shelter, Central crash barrier, Near/Offside crash barrier, Submerged in water, Entered ditch, Other permanent object, Wall or fence, or Data missing or out of range.

- 1st Point of Impact
    - Indicates the first point of impact as Did not impact, Front, Back, Offside, Nearside, or Data missing or out of range.

- Was Vehicle Left Hand Drive
    - Indicates if the vehicle was left-hand drive as No, Yes, or Data missing or out of range.

- Journey Purpose of Driver
    - Indicates the journey purpose of the driver as Journey as part of work, Commuting to/from work, Taking pupil to/from school, Pupil riding to/from school, Other, Not known, Other/Not known (2005-10), or Data missing or out of range.

- Sex of Driver
    - Indicates the sex of the driver as Male, Female, Not known, or Data missing or out of range.

- Age of Driver [Only exists for 2016, 2015, & 2014]
    - Indicates the age of the driver as a integer.

- Age Band of Driver
    - Indicates the age band of the driver as 0 - 5, 6 - 10, 11 - 15, 16 - 20, 21 - 25, 26 - 35, 36 - 45, 46 - 55, 56 - 65, 66 - 75, Over 75, or Data missing or out of range.

- Engine Capacity (CC)
    - Indicates the engine capacity as an integer.

- Vehicle Propulsion Code
    - Indicates the propulsion of the vehicle as Petrol, Heavy oil, Electric, Steam, Gas, Petrol/Gas (LPG), Gas/Bi-fuel, Hybrid electric, Gas Diesel, New fuel technology, Fuel cells, Electric diesel, or Undefined.

- Age of Vehicle (manufacture)
    - Indicates the age of the vehicle as an integer.

- Driver IMD Decile
    - Indicate the driver IMD decile as Most deprived 10%, More deprived 10-20%, More deprived 20-30%, More deprived 30-40%, More deprived 40-50%, Less deprived 40-50%, Less deprived 30-40%, Less deprived 20-30%, Less deprived 10-20%, Least deprived 10%, or Data missing or out of range.

- Vehicle IMD Decile [Only exists for 2016, & 2015]
    - Indicate the vehicle IMD decile as Most deprived 10%, More deprived 10-20%, More deprived 20-30%, More deprived 30-40%, More deprived 40-50%, Less deprived 40-50%, Less deprived 30-40%, Less deprived 20-30%, Less deprived 10-20%, Least deprived 10%, or Data missing or out of range.

- Driver Home Area Type
    - Indicate the driver home area type as Urban area, Small town, Rural, or Data missing or out of range.

The data was merged and / or concatenated into various other DataFrames throughout the investigation. These DataFrames were saved as .csv files and placed into the …\EditedDataFiles directory.

Summary Statistics

The majority of the data fields are categorical. Of the numerical fields, summary statistics were calculated where relevant[1]:

Data Table	Year	Column	Count	MIN	25%	MEDIAN	75%	MAX	MEAN	STD
Accidents	2016	Number of Vehicles	136621	1	1	2	2	16	1.848179	0.710117
		Number of Casualties	136621	1	1	1	1	58	1.327644	0.789296
	2015	Number of Vehicles	140056	1	1	2	2	37	1.841014	0.710046
		Number of Casualties	140056	1	1	1	1	38	1.329390	0.795427
	2014	Number of Vehicles	146322	1	1	2	2	21	1.835179	0.700208
		Number of Casualties	146322	1	1	1	1	93	1.329103	0.857469
	2013	Number of Vehicles	138660	1	1	2	2	67	1.823980	0.726114
		Number of Casualties	138660	1	1	1	1	70	1.324607	0.801197
	2012	Number of Vehicles	145571	1	1	2	2	18	1.826442	0.703703
		Number of Casualties	145571	1	1	1	1	42	1.344519	0.805668
	2011	Number of Vehicles	151474	1	1	2	2	34	1.823118	0.710169
		Number of Casualties	151474	1	1	1	1	87	1.346436	0.856727
	2010	Number of Vehicles	154414	1	1	2	2	19	1.822380	0.704551
		Number of Casualties	154414	1	1	1	1	43	1.351225	0.823076
	2009	Number of Vehicles	163554	1	1	2	2	32	1.826229	0.716708
		Number of Casualties	163554	1	1	2	2	48	1.358243	0.813896

Data Table	Year	Column	Count	MIN	25%	MEDIAN	75%	MAX	MEAN	STD
Casualties	2016	Age of Casualty	178538	0	22	33	50	101	36.928251	18.822219
	2015	Age of Casualty	183195	0	22	33	49	104	36.700259	18.690425
	2014	Age of Casualty	191403	0	22	33	49	102	36.67418	18.70622

Data Table	Year	Column	Count	MIN	25%	MEDIAN	75%	MAX	MEAN	STD
Vehicles	2016	Age of Driver	223082	1	27	38	51	101	40.287450	16.287479
		Engine Capacity (CC)	193720	11	1248	1598	1995	91000	1868.633998	1631.050907
		Age of Vehicle	184325	1	4	8	12	75	8.138050	5.154017
	2015	Age of Driver	228017	1	1	1	2	37	40.169066	16.152652
		Engine Capacity (CC)	185556	13	1248	1598	1995	22311	1896.149206	1677.424281
		Age of Vehicle	176075	1	4	8	11	105	8.011359	4.995217
	2014	Age of Driver	239075	1	27	38	51	100	40.303327	16.190651
		Engine Capacity (CC)	202287	2	1248	1598	1995	17696	1895.110002	1661.106155
		Age of Vehicle	193178	1	4	8	11	80	8.043877	4.853613
	2013	Engine Capacity (CC)	191487	4	1275	1598	1995	25422	1902.391196	1661.680529
		Age of Vehicle	183187	1	4	8	11	111	7.919132	4.733979
	2012	Engine Capacity (CC)	198957	8	1298	1598	1995	91000	1890.679081	1640.990127
		Age of Vehicle	191241	1	4	8	11	82	7.775791	4.577431
	2011	Engine Capacity (CC)	208102	4	1298	1598	1995	61000	1903.987045	1651.937390
		Age of Vehicle	200779	1	4	7	10	99	7.586695	4.529736
	2010	Engine Capacity (CC)	216591	11	1299	1598	1995	91000	1935.082538	1768.470903
		Age of Vehicle	209472	1	4	7	10	110	7.388443	4.449240
	2009	Engine Capacity (CC)	230231	1	1299	1598	1995	99999	1899.310023	1643.936015
		Age of Vehicle	222876	1	4	7	10	95	7.160802	4.422070

Looking at the quartile spread, the mean, and the standard deviation for the above fields, there is little to no discernible difference between the years 2016 and 2009. The overall number of accidents appears to have fallen though, which is nice to hear.

[1] Calculation of a mapping coordinate such as longitude makes very little sense. Equally whilst the speed limit field is an integer, it is - fundamentally - categorical.

Graphical EDA

Univariate

The first purpose of the analysis tasked me with summarizing the main characteristics of the data, and to obtain interesting facts that are worth highlighting. To perform this task, seaborns countplot() was used to plot the count of each object in each column and use the hue argument to colour-coordinate by year. The plots are available in the …\Images\AllFieldsByYear directory. Looking at the plots, the key points of information are noted in the bulleted points below:

- **1st_Point_of_Impact**
  - In order of frequency:
    - Front, Back, Offside, Nearside, & Did not impact
  - Everything appears the same from year to year.

- **1st_Road_Class**
  - In order of frequency:
    - A, Unclassified, B, C, Motorway, & A(M)
  - There appears to be a decreasing trend for A, B, & C. There appears to be an increasing trend for Unclassified. Motorway and A(M) appear to be the same year to year.

- **2nd_Road_Class**
  - In order of frequency:
    - NaN, Unclassified, A, C, B, Motorway, & A(M).
  - There appears to bea downward trend on Unclassified, A, C, and B. Motorway appears to be the same. One can't tell with A(M).

- **Accident_Severity**
  - In order of frequency:
    - Slight, Serious,& Fatal.
  - There appears to be a downward trend in Slight. Severe and Fatal appear to be the same.

- **Age_Band_of_Casualty**
  - In order of frequency (Top 5):
    - 26 - 35, 36 - 45, 46 - 55, 21 - 25, & 16 - 20
  - There appears to be a downward trend in the 16 - 20, 21 - 25, & 36 - 45 age bands. There appears to be an upwards trend in the 26 - 36 age band. The 21 - 25 and 46 - 45 age bands appear to be the same.

- **Age_Band_of_Driver**
  - In order of frequency (Top 5):
    - 26 - 35, 36 - 45, 46 - 55, 21 - 25, & 56 - 65
  - There appears to be a downward trend in the 21 - 25, & 36 - 45 age bands. The 26 - 35, 46 - 55, and 56 - 65 age bands appear to be unchanged.
  - This is particularly interesting as it provides evidence against the common trope that the 16 - 20 and 21 - 25 age bands are the "young and reckless boy-racers".

- **Bus_or_Coach_Passenger**
  - In order of frequency:

        - Not a bus or coach passenger (dominates), Seated passenger, Standing passenger, Boarding, Alighting, & NaN.
    - There appears to be a downward trend in Seated passengers. The other fields appear to be the same.

- **Car_Passenger**
    - In order of frequency:
        - Not a car passenger, Front seat passenger, rear seat passenger, & NaN.
    - All the fields appear to be the same year on year.

- **Carriageway_Hazards**
    - In order of frequency:
        - None (dominates), Other object in road, Any animal in carriageway (except riden horse), Pedestrian in carriageway, vehicle load on road, Previous accident, and NaN.
    - There appears to be a downward trend in None, Other object in road, and Any animal in carriageway (except ridden horse). Pedestrian in carriageway seemed to spike from 2012 to 2014 and fall again to 2016. NaN data appears to be climbing rapidly. Previous accidents approximately doubled in 2015.

- **Casualty_Class**
    - In order of frequency:
        - Driver or rider, Passenger, and Pedestrian.
    - They all appear to be the same year to year.

- **Casualty_Home_Area_Type**
    - In order of frequency:
        - Urban area, NaN - Rural, and Small town.
    - They all appear to be the same year to year.

- **Casualty_IMD_Decile (Data only available for 2015 and 2016)**
    - In order of frequency (Top 5):
        - Nan, Most deprived 10%, Most deprived 10 - 20%, Most deprived 20 - 30%, and Most deprived 30 - 40%.
    - The data appears to be falling year to year, but given the NaN increased and there are only two data points, on cannot make a reasonable judgement.

- **Casualty_Severity**
    - In order of frequency:
        - Slight, Serious, and Fatal.
    - They appear to be the same year to year.

- **Casualty_Type**
    - In order of frequency (Top 5):
        - Car occupant (dominates), Cyclist, Pedestrian, Motorcycle under 125CC and under rider or passenger, Motorcycle over 500CC rider or passenger.
    - Cyclist accidents saw a spike in 2014. Motorcycle under 125CC and under rider or passenger appear to be rising.

- **Day_of_Week**
  - In order of frequency:
    - Friday, Monday - Thursday, Saturday, and Sunday.
  - They all appear to show a downward trend.

- **Did_Police_Officer_Attend_Scene_of_Accident**
  - In order of frequency:
    - Yes, No, No (Self Report Only).
  - The Yes appears to be showing a downward trend. The No appears to be the same.

- **Driver_Home_Area_Type**
  - In order of frequency:
    - Urban area, NaN, Rural, Small Town.
  - They all appear to be the same year to year.

- **Driver_IMD_Decile**
  - In order of frequency:
    - NaN (dominates), Most Deprived 10% - Most deprived 40 - 50%.
  - The data appears to be missing for 2014.
  - From the Most deprived 40 - 50% to the least deprived 10%, there appears to be a downward trend. From the Most deprived 10% to the Most deprived 40 - 50% there appears to be no change year to year. This would imply that the wealthier drivers are having fewer accidents.

- **Hit_Object_in_Carriageway**
  - In order of frequency (Top 5):
    - None (dominates), Kerb, Parked vehicle, Bollard or refuge, and Other object.
  - Missing data seems to have drastically risen. Parked vehicles appear to be increasing. Kerb appears to have peaked in 2014 - 2015.

- **Hit_Object_off_Carriageway**
  - In order of frequency (Top 5):
    - None (dominates), Other permanent object, Tree, Wall or fence, Entered ditch - Road sign or traffic signal.
  - They all appear to be declining and Other permanent objects has drastically fallen. Wall or fence has drastically risen.

- **Journey_Purpose_of_Driver**
  - In order of frequency:
    - Not known (dominates), Journey as part of work, Commuting to/from work, Other, Taking pupil to/from school, Pupil riding to/from school.
  - Journey as part of work appears to be falling. Other saw large spikes in 2012 and 2016.

- **Junction_Control**
  - In order of frequency:
    - Giveway or uncontrolled, NaN, Auto-traffic signal, Stop sign, Authorised person, Not a junction or within 20 metres.
  - Giveway or uncontrolled appears to be decreasing. The rest appear unchanged.

- **Junction_Detail**
  - In order of frequency (Top 5):
    - Not a junction or within 20 metres, T or staggered junction, Crossroads, Roundabout, Private Drive or entrance.
  - T or staggered junction appears to be decreasing.

- **Junction_Location**
  - In order of frequency (Top 5):
    - Not a junction or within 20 metres of a junction, Approaching junction or waiting / parked at junction approach - Mid junction (on roundabout or on main road), Cleared junction or waiting / parked at junction/exit.

- **Light_Conditions**
  - In order of frequency:
    - Daylight (dominates), Darkness lights lit, Darkness no lighting, Lighting unknown, Darkness lights unlit, NaN
  - Daylight appears to be falling. The rest appear to be the same year to year.

- **Pedestrian_Crossing-Human_Control**
  - In order of frequency:
    - None within 50 metres (dominates), Control by other authorised person, Control by school crossing, NaN.
  - There appears to be a lot of variation in Control by other authorised person.

- **Pedestrian_Crossing-Physical_Facilities**
  - In order of frequency (Top 5):
    - No physical crossing facilities within 50 metres (dominates), Pedestrian phase at traffic signal junction, Pelican, puffin, toucan, or similar non-junction pedestrian light crossing, Zebra, and Central refuge.
  - Pelican, puffin, toucan, or similar non-junction pedestrian light crossing appear to be decreasing.

- **Pedestrian_Location**
  - In order of frequency (Top 5):
    - Not a pedestrian (dominates), In carriageway, crossing elsewhere, Crossing on pedestrian crossing facility, and In carriageway not crossing.

- **Pedestrian_Movement**
  - In order of frequency (Top 4):
    - Not a pedestrian (dominates), Crossing from drivers nearside, Crossing from drivers offside, Unknown or Other.

- **Pedestrian_Road_Maintenance_Worker**
  - In order of frequency:
    - No/Not applicable (dominates), Not known, Yes.

- **Police_Force**
  - In order of frequency (Top 5):

    - Metropolitan Police (dominates), WestMidlands, West Yorkshire, Thames Valley, and Kent.
  - Kent appears to have had a spike in 2013.
  - Greater London would appear to dominate due to large traffic in and out of the capital. The next most frequent being the West Midlands and West Yorkshire may indicate the winding roads are a factor.

- **Propulsion_Code**
  - In order of frequency (Top 5):
    - Petrol (dominate), Heavy Oil (dominate), Undefined (dominate), Hybrid electric, Gas/Bi-fuel.
  - Hybrid electric is seeing a stark increase. Electric and Heavy oil are also seeing an increase. Petrol is seeing a downward trend alongside Gas/Bi-Fuel and Petrol/Gas (LPG).

- **Road_Surface_Conditions**
  - In order of frequency:
    - Dry, Wet or damp, Frost or Ice, Snow, Flood over 3m deep, and NaN.
  - I would have thought that environmental conditions made up more accidents, but in fact dry conditions are the most frequent. I hypothesise that when the conditions are bad, people take more care (drive slower) or perhaps don't even drive and postpone the journey. Additionally more Dry days compared to Wet or Damp days or Snow days. This could perhaps be investigated further for the number of accidents relative to the total amount of days over which that weather condition was active.

- **Road_Type**
  - In order of frequency:
    - Single Carriageway, Dual carriageway, Roundabout, One-way Street, slip road, unknown.

- **Sex_of_Casualty**
  - In order of frequency:
    - Male, Female.
  - Men make up ~20% more casualties than women.

- **Sex_of_Driver**
  - In order of frequency:
    - Male, Female, Unknown.
  - Men make up ~50% more of the drivers than women. This goes to show that the stereotype of "women are terrible drivers" is in fact incorrect.

- **Skidding_and_Overturning**
  - In order of frequency:
    - None (dominates), Skidded, & Skidded and Overturned - Overturned.
  - Skidded appears to be declining.

- **Special_Conditions_at_Site**
  - In order of frequency:
    - None (dominates), Roadworks, Oil or Diesel - Mud, & Auto-Traffic signal out - Road surface defective.

- **Speed_limit**
    - In order of frequency:
        - 30 (dominates), 60, 70, 40, 50.
    - Accidents at 30 appear to be decreasing.

- **Towing_and_Articulation**
    - In order of frequency:
        - No tow/articulation (dominates), Articulated vehicle, Single trailer, Other tow, Caravan, Double or multiple trailer.
    - The missing data appears to be drastically increasing.

- **Urban_or_Rural_Area**
    - In order of frequency:
        - Urban, and Rural.

- **Vehicle_Leaving_Carriageway**
    - In order of frequency (Top 5):
        - Did not leave carriageway (dominates), Nearside, Offside, Nearside and rebounded, and Offside onto central reservation.
    - Nearside and Offside appear to be falling.

- **Vehicle_Location-Restricted_Lane**
    - In order of frequency:
        - On main carriageway - not in restricted lane (dominates), Footway (pavement), Bus lane, On lay-by or hard shoulder, and Cycle lane (on main carriageway).
    - The amount of missing data appears to be drastically increasing.

- **Vehicle_Manoeuvre**
    - In order of frequency:
        - Going ahead (other) (dominates), Turning right, slowing or stopping, Waiting to go / held up, Going ahead (right-hand bend).
    - There is notable variation here year to year.

- **Vehicle_Type**
    - In order of frequency:
        - car (dominates), Pedal cycle, Van / Goods 3.5 tonnes mgw or under, Motorcycle 125CC or under, and Motorcycle over 500CC.
    - Pedal cycles seemed to have a spike in 2014. Goods vehicle (unknown weight) appears to be increasing.

- **Was_Vehicle_Left_Hand_Drive?**
    - In order of frequency:
        - No (dominates), and Yes.

- Weather_Conditions
    - In order of frequency:
        - Fine no high winds (dominates), Raining no high winds, the rest appear equal, and Snowing + high winds last.
    - Raining no high winds appears to have spikes in 2012 and 2014.

To simplify, the most common accident is a “slight” one and occurs between a single male driver aged between 26 - 35 years from an urban area and a 26 - 35 male casualty inside of another car; the casualty is also from an urban area. The driver is in a right-hand controlled petrol car on a single carriageway road with a 30 mph speed limit. The road conditions are dry, it is daylight, and the weather is fine with no high winds. The collision occurs on a Friday and is a front end collision. It occurred whilst the car was travelling straight ahead. The journey purpose of the driver was unknown. The police attended the accident. It is more likely that it didn’t occur near a junction (20m), but if it did occur at a junction, it occurred at a give-way or a controlled junction.

Bivariate

The second purpose of the analysis tasked me with identifying and quantifying associations (if any) between the number of causalities (in the Accidents table) and other variables in the data set. To investigate this, seaborn’s countplot() was used again, but this time within seaborn’s factorplot(). This allowed the number of casualties to be plotted with the hue argument separating the data by each of the other fields and the col argument allowing the year of the data to separate it further. The plots can be found in the …\Images\ NumOfCas_Banded_ByAllFields_ByYear directory.

To plot the number of casualties, a new field was created to band them together. The bands chosen were: 1, 2, 3, 4, 5, 6-10, 11-20, and 21+.

The plots - I.M.O. - are visually noisy due to separation by three fields. The patterns of the data, however, appear to be undiscernible on year to year, therefore moving forward, and for the Modelling chapter later, only the 2016 data will be used. Using a reduced (only 2016) DataFrame, the Number of Casualties was plotted against every other field in the DataFrame using the hue argument. The plots were saved in the …\Images\ NumOfCas Vs AllFields 2016 directory.

Inspecting the plots, three relationships stood out with a potential fourth.

1. When hued with the **Pedestrian_Crossing-Human_Control** field, the “Control by other authorised person” category falls sequentially from 1 to 5 and then rises again in 6-10 and rises again further to the 5 level in 11-20 and 21+. To me, this implies that when an authorised person is controlling the crossing, it is because it is particularly dangerous or it needs to be synchronised with another form of transport, a train perhaps. Therefore the increase would imply that the riskier situation has more potential to go badly wrong.
2. When hued with the **Special_Conditions_at_Site** field, the “Auto-traffic signal out” category had no casualties for 5 or 6-10 but did for 11-20. To me, this implies that a junction or crossing of some description has the potential to cause multiple accidents if faulty. The “Mud” field also showed an increase in the 6-10 band compared to 4 and 5.
3. When hued with the **Rural_or_Urban_Area** field, the absolute count of “Urban” is greater than “Rural” for casualties 1, 2, and 3, but for casualties 4, 5, 6-10, 11-20, and 21+ the categories swap. This implies that you are more likely to see higher number of casualties in “Rural” areas compared to “Urban” areas.
4. When hued with the **Pedestrian_Movement** field, there was an increase in the “Unknown or Other” category from 5 to 6-10 and again to 11-20. Given the lack of information, I am unable to discern anything from this.

Given the col argument is unused in the factorplot(), it would be possible to explore tertiary relationships. This was not done due to time restrictions, but it is noted as a possibility for further EDA.

Modelling & Evaluation

The third purpose of the analysis tasked me with exploring whether it is possible to predict accident hotspots based on the data. I will first explain how I defined an accident "hotspot", how I implemented it using Python, and a critical discussion of it's approach.

My interpretation of a "hotspot" is frequency, not severity. Simply having a casualty at an incident in a given area is not enough, having a casualty is not important for that matter, what is important is the amount of incidents that occur in a given area. Secondly, in defining a hotspot, I don't believe that it should be a collection of circumstances, but it should be a collection of environmental factors. I.E. it shouldn't matter if you are a man from the least 40-50% deprived region, but it should matter if it is raining at a junction and you are towing a trailer.

To implement my interpretation of a hotspot, I used a 2016 only DataFrame containing the information of Accidents, Casualties, and Vehicles, I then reduced the columns down to simply the **Date** and the **Local_Authority_(District)**. I converted the date column (stored as a string) to a datetime object and set this as a DateTimeIndex. Grouping by the date index and the **Local_Authority_(District)** column, I counted the occurrence of each district per day, the resulting DataFrame went from looking something resembling:

```
                Date            Local_Authority_(District)
Datetime
2016-11-01      01/11/2016      Brent
2016-11-01      01/11/2016      Brent
2016-11-01      01/11/2016      Bexley
2016-11-01      01/11/2016      Hillingdon
2016-11-01      01/11/2016      Merton
```

to something resembling:

```
                        Local_Authority_(District)      Counts
Datetime
2016-01-01              Westminster                     4
2016-01-02              Westminster                     1
2016-01-03              Westminster                     2
2016-01-04              Westminster                     6
2016-01-05              Westminster                     2
```

when restricted to a single district. I then looped through all of the 416 districts, and for each district, search for all of them which experienced at least 1 accident a day for 6 out of 7 days a week (312 days out of 365 days a year) of the year. This returned the following districts as "hotspots":

- Westminster (356),
- Camden (321),
- Hackney (335),

- Tower Hamlets (345),
- Lewisham (330),
- Southwark (333),
- Lambeth (354),
- Wandsworth (336),
- Kensington and Chelsea (313),
- Redbridge (312),
- Newham (333),
- Bromley (321),
- Croydon (323),
- Hounslow (335),
- Ealing (339),
- Brent (326),
- Barnet (334),
- Haringey (326),
- Enfield (326),
- Liverpool (344),
- Cheshire East (325),
- Cheshire West and Chester (315),
- Bradford (355),
- Kirklees (327),
- Leeds (363),
- Wakefield (312),
- Doncaster (322),
- Sheffield (348),
- Kingston upon Hull, City of (322),
- East Riding of Yorkshire (322),
- Birmingham (366),
- Nottingham (322),
- Leicester (320),
- Brighton and Hove (317),
- Cornwall (354),
- Bristol, City of (345),
- Wiltshire (338),
- Edinburgh, City of (337), &
- Glasgow City (344).

The code used to generate these results was `HotSpotTest.py`.

Discussion

The most obvious criticism of this model is that it does not use any of the data from the tables, in its implementation, it merely calculates the frequency of an accident every day and then looks for the districts where at least one accident occurs 6 days a week on average. By my interpretation, this is a hotspot, but it could be improved upon had I not run out of time. Ideas for improvement could include:

- It is currently not independent of traffic flow, therefore larger districts, which see larger traffic flow - whether that be industrial or domestic - would of course have more accidents per day. This could be controlled with more information, perhaps the traffic rate of a district, or its population density of registered drivers. Then the number of accidents per driver could be calculated and used as the metric.

- It is currently grouping all accidents as an “incident”. This could be split and perhaps separate out only “slight” or only “serious” accidents. Furthermore, certain environmental features could be selected for certain road features. I.E. what are the proportion of accidents which occur in a district for a specific type or collection of types of accidents making them “hotspots” for those types of accidents. Leading on from this idea, can one hold constant environmental conditions, I.E. in a certain district, all of their accidents which occur in the rain occur at T-Junctions.

Deployment & Areas of Future Work

Interactive plotting with Bokeh

I intended to make use of, Bokeh to make the plots and other analysis steps implemented in this report interactive such that one could select from a drop-down list the x, the hue and the col arguments amongst other tools, however I unfortunately ran out of time to implement this.

GeoMapping with Basemap

Having generated a list of hotspots, they have associated with them a series of latitudes and longitudes defining their boundaries. I planned to plot these regions to visually demonstrate the regions. Additionally, I intended to plot (similar to Google drop pins) the locations of the accidents themselves. These accidents could then be filtered and reduced to show only the accidents involving roundabouts, for example.

I didn’t run out of time to implement this, but I instead encountered technical difficulties. I use PyCharm as my IDE, which downloads its libraries from PIP which no longer hosts Basemap due to the size. Researching on the forums, it is possible to install Basemap directly, but you also need to do other reconfigurations in the terminal. A second option is to use the IDE Anaconda as conda does host Basemap. My reasoning against both of these options is that whilst I may get it to work, I run the very real risk of messing something up, and given the importance of working on this project and the limited time available, I opted to leave this for the future.

About the Author

Hello, World! If you wish to contact me, please select from the following options:

E-Mail Address and Personal Website

E-Mail:	Matthew.William.Noble@gmail.com
Personal website:	http://matthewnoble.info/index.html

CV and Coding Portfolio

CV (pdf):	https://drive.google.com/open?id=1d6mdd39LRKL2DLfNmaY6xiOlIBJm5653
CV (online):	http://matthewnoble.info/MyCV.html
GitHub:	https://github.com/MatthewWilliamNoble/CodingPortfolio

Social Media

Facebook:	https://www.facebook.com/matthew.w.noble
LinkedIn:	https://www.linkedin.com/in/matthew-w-noble/
Twitter:	https://twitter.com/Matthew_W_Noble
Instagram:	https://www.instagram.com/matthew.w.noble/
Kaggle:	https://www.hackerearth.com/@matthew104
HackerEarth:	https://www.hackerearth.com/@matthew104
HackerRank:	https://www.hackerrank.com/matthew_noble_11